DOCUMENT RESUME

ED 290 041 CE 049 426

AUTHOR Cantor, Jeffrey A.; Walker, C. Lee
TITLE The Design of Criterion-Referenced Tests to Support Vocational Program Evaluation.
PUB DATE Dec 87
NOTE 12p.; Paper presented at the American Vocational Association Convention (Las Vegas, NV, December 6-10, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Content Validity; *Criterion Referenced Tests; Evaluation Methods; Postsecondary Education; *Program Evaluation; Secondary Education; *Test Construction; *Test Use; *Test Validity; *Vocational Education

ABSTRACT

This paper discusses and describes an industry-based process for the design and development of criterion-referenced tests in vocational-technical subject areas to support vocational program evaluations. The following topics are covered in the paper: (1) use of criterion-referenced tests; (2) test design; (3) preparation of test materials; and (4) validation of test materials. Guidelines are provided in each of these areas for the reader to follow in developing and validating criterion-referenced tests. (KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED290041

THE DESIGN OF CRITERION-REFERENCED TESTS TO SUPPORT VOCATIONAL PROGRAM EVALUATION

by

Jeffrey A. Cantor
Director of Training

and

C. Lee Walker
Manager, Testing Programs

DDL OMNI Engineering Corp.
Chantilly, Virginia

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J Cantor

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

paper presentation
Annual Meeting
American Vocational Association
Las Vegas, Nevada
December, 1987

CE049426

THE DESIGN OF CRITERION-REFERENCED TESTS TO SUPPORT VOCATIONAL PROGRAM EVALUATION

Jeffrey A. Cantor (Director of Training)
and
C. Lee Walker (Manager, Testing Programs)
DDL OMNI Engineering Corp.

The purpose of this presentation is to discuss and describe an industry developed process for the design and development of criterion-referenced tests in vocational-technical subject areas, to support vocational program evaluations. The illustrations which follow this discussion have been designed to facilitate the test development process and can be adapted to individual project needs.

DISCUSSION

Criterion-referenced (C-R) testing in vocational-technical education and training has become an area of interest primarily as a result of the movement towards accountability for learning outcomes during the period of the 1970's. Increased emphasis was placed upon C-R testing in voc nal training programs as witnessed in regulations promulgated for program funding in the Job Training Partnership Act and the Perkins Act. To address the requirements for sound development of C-R tests, program developers and planners turned their attention to the body of knowledge developed by testing and measurement reasearchers.

A requirement for C-R tests as part of a program of vocational trades training was addressed in the U.S. Navy's shipyard community. Based on the principles and practices of criterion-referenced testing and measurement, a process was devised to develop content valid tests in several trades practiced within the shipyard.

These tests were developed in concert with shipyard training personnel and involved significant input from senior trades technicians. Both the development of the tests and the design of C-R scores involve planned workshops. This presentation will expound upon, and display these workshops processes and theories, and how they can offer vocational educators a tool for C-R test development.

The Naval Shipyards have implemented a Shipyard Training Modernization Program in order to effect the efficiency and standardization which can be achieved through sound practices in training material development and training management. A component of this program is a system for assessing the skills and performance of trades personnel using objective measures. This system will be used to verify the attainment of specific skills and to diagnose areas where either individual or system training can be improved. In particular, objective measurement is viewed as an appropriate tool in:

a. Monitoring the progress of apprentices through their formal apprenticeship training.

b. Placement of personnel entering shipyard employment with previous related experience.

c. Verifying the status of helpers and limited workers being considered for advancement.

d. Analyzing training with respect to effectiveness and the knowledge retained by individual workers.

Whenever tests are used in Naval Shipyards in a way which may effect personnel decisions, the tests must be job related. Job related tests are also known as Content Valid tests. Guidance has been provided by the courts and legal community, the Equal Opportunity Commission (EEOC), and professional testing associations on writing of content valid tests. The process described in this paper is aimed at enabling Naval shipyard personnel to write content valid tests, and to maintain appropriate records for addressing any challenges to the validity of the tests.

TEST DEVELOPMENT

The test development process has three distinct phases: (1) test design, a determination of what is to be tested, (2) preparation of the test materials, and (3) validation of the test materials. Each phase has its own requirements and records which together provide an audit trail for content validity. Within each phase, there is flexibility of method to suit constraints of resources and time.

TEST DESIGN. Test design specifies what is to be tested. Test design is viewed in three parts. First is a definition of who is to be tested and why. Second, is an analysis of the tasks, knowledges, skills and abilities which comprise the job or position to be measured. Third is the selection of measurement points and measurement methodology. Collectively, these parts build a specification from which the test is written. This "front end" analysis is of prime importance to establishing the content validity of a test. Each test is related to a job or task analysis. The sources of the analyses can vary. For example, the tests for Inside Machinists and the Electrical Trades used a Portsmouth Naval Shipyard (PNSY) analysis called a "Prime Line - Map Out". The Marine Machinist test incorporated an ad hoc analysis of the task to be measured. The "Secondary System Cleanliness Test" used a "retention" test based on a course of instruction, and used the course learning objectives as the basis of the test.

Purpose of the Test. The purpose of the test, (e.g., who is to be tested and how the resulting measurement information is to be used) must be decided upon before any actual development of measurement instruments begins. A form for establishing a plan for test development was devised. This plan is a working document which should be completed, reviewed and discussed early in the development process, and then revised as required to reflect the test.

Completion of the Test Plan. The Test Plan should be completed by the person who is asking for the test, or the person responsible for writing the test. It is imperative that all people involved have a common understanding of the plan before further development takes place. Information gathered should include:

a. Persons to be tested.

b. Purpose of the test.

c. Planned Test Method. Ideally the best measurement approach is determined by an analysis of the purpose of the assessment and the behavior to be measured. Administrative time or resource constraints may dictate whether written or performance testing is to be used. If the planned testing method is constrained, or a prefered test method is known, it should be so stated.

d. Scoring Reference. Tests will normally be either of a pass/fail nature, where measurement is judged based upon an established standard or "criterion"; or will be of a diagnostic nature in which the measurement will show a relationship between each person's mastery/competence to each other. This second type of scoring reference is called normative.

e. Approximate Test Time. Constraints on the time available for testing as well as a desired test length should be noted in the plan. Consideration should be given to the test purpose and subject matter to be covered.

f. Subject Area(s). The subject area(s) for the test should be decided upon by the person requesting the test; or alternatively by the persons writing the test. The number of items dedicated to each subject area will normally be decided upon and noted as the test is developed.

Task Analysis. The content or subject areas appropriate to the test must be analyzed in detail to determine the tasks, knwoleges, skills and abilities which are critical to carrying out the work associated with each of the areas. Subsequent paragraphs will discuss some forms and options for developing and recording this information. However, important concepts to understand are: (1) Critical measurement points should be reviewed and agreed upon by several persons whos experience qualifies them to make the judgement and (2) Level of detail in the analysis need only be deep enough to satisfy the intended purpose of the test.

Participant Background Information. A form was devised for recording relevant work and training experience about persons who are asked to make judgements on matters concerning test content. There is no specific guidance on how much experience is enough, but persons should have had recent and relevant experience on the task being considered. A mix of persons who supervise the work and who actually do the work is desirable.

Task Analysis Record. Figure 2 is a set of forms for recording judgements on task analysis data regarding the degree of perceived importance of the various components of the task analysis data to the overall purpose of the test. Variables include (1) Duties, (2) Tasks, (3) Task elements and (4),

Knowledges, skills and abilities. For example, if it is desirable to measure performance of a specific job for its own sake, (e.g., qualification to overhaul a specific valve), then tasks and task elements will probably be the end product of the analysis. If the measurement is in order to generalize the results, then one will identify the knowledges, skills and abilities which constitute performance. In this last case the breakdown of duties into tasks and into task elements need proceed only far enough so that one can be sure the required knowledge, skills and abilities can be identified.

Measurement Points and Methodology. The task analysis serves to identify critical elements in the area to be measured. Those elements must be reviewed to determine (1) the element that should be measured, (2) whether written or performance testing is required or acceptable for the measurement, (3) how many questions or trials are needed to establish proficiency on the element. This first level of decision must be followed up by considerations of the time required to test and, if performance testing is preferred, the availability of appropriate tools, machines and materials. In addition, consideration is required on whether drawings and documents will be used to support the test or whether measurement will be carried out "closed book." These decisions should be recorded on the task or knowlege, skills, ability listing in as much detail as required to make the specifications for the test clear to the person who will write the test. Although this process of definition and review is time consuming it is important to ensure "content validity" of the test. It also makes the work of the test writer much easier.

Task Analysis Review. If the task analysis and measurement selection is carried out by one or two people, for instance, trade instructors, the resulting task analysis and selected measurement points should be reviewed by other persons actively working in the subject area. If the analysis has been carried out by a working group of competent persons specifically gathered for the purpose there is no requirement for external review.

PREPARATION OF TEST MATERIALS. Testing materials will always include a protor guide and the test. In addition the testing material may include answer sheets, examinee directions, examinee informatio sheets, supporting documents and drawings, tools, materials, etc. Tests generally will be either written tests or performance tests, although in some instances a test may contain elements of both written and performance testing.

Written Tests. A specification has been developed for preparing written tests. A test item form is provided which should be completed for each discrete question. The form is designed to facilitate multiple choice item design, but may be locally modified to accommodate other written question types.

Performance Tests. A specification has also been developed for preparing performance tests materials where a determination of the acceptability of the examinees performance will be made by observation of work in progress and/ or by examination of a product.

Combination Tests. Tests combining written portions and performance portions may be developed so long as the directions in the proctor guide are clear and easy to follow. It is also important that a method for combining or presenting the scores be provided.

Examinee Directions. These are specifically written directions to the examinee which provide information on procedures, materials, constraints, safety and scoring. Written directions are always included with performance tests and may be included with written tests at the option of the person responsible for administering the test.

Examinee Information Sheet. Analysis of test results in many cases is more meaningful when information with resepect to the examinees training background and work experience is available. This is particularly true when the test population is either small or consists of persons with specialization within the trade or population.

Answer Sheets. Answer sheets as an alternative to marking directly on the test provide an efficient means of scoring and of maintaining records on file for analysis. Marking directly on the test may be preferable when people are tested as individuals or in small groups and the test results are to be discussed with the examinee.

Supporting Materials. All materials required to support a test will be listed in the Proctor Guide.

References for Measurement Points. Where a test question or performance check point is based on a document, drawing or directive governing trade work, the document, drawing or directive should be cited on the test item form or check list as appropriate. Questions which are matters of common trade practice and are not addressed in shipyard paper work should state

"CTP" in the reference block on the test item form. It is not necessary to annotate each performance check point unless some specific standard is stated e.g., tighten nut to 57 ft-lbs (CTP).

VALIDATION OF TEST MATERIALS. Validation of test materials provides two valuable activities. First is a verification that the test "as written" measures critical elements of the areas being tested. Secondly, validation ensures that the test as a measuring device measures fairly and consistently. The process of validation incorporates a review of the testing material and administration of the material on a trial basis.

Content Validation. Content validation is accomplished by using a sample of 5 to 10 different persons from those who participated in the development of the test. These persons must include representatives of the group to be tested but may also include supervisory personnel. The verification is best conducted in a group setting.

In the verification each person takes the test and then judges individual items with respect to difficulty and with respect to the items importance to the area being measured. Forms and directions were devised for carrying out the verification. This same process may be extended to provide a "first cut" at pass/fail standard setting on criterion tests. In addition to the Importance/Difficulty judgements derived from the verification, the test results should be reviewed for indications of problem areas. Overall scores should be calculated and the results of each question or check point should be looked at for high and low success rates. Wrong answers consistently chosen should be examined for the possibility of mis-keying. Problem questions should be discussed with the participants. The participant should note problem areas in the test on their copies of the test to assist in review. A Participant Background Information form should be completed by each member of the group and retained in an audit file on the test.

Pilot Administration of the Testing Materials. The testing materials are to be piloted by small groups of individuals different from those who had done the test development. The individuals selected for the pilot testing represent the population for whom the test is designed. The tests are scored at the time they are administered and comments are solicited from the pilot population on individual test readability and appropriateness items. Statistics and

response patterns on item alternatives are calculated for each test, as appropriate to shipyard needs.

Statistical Validation. Statistical validation is a must for tests which are to be used for personnel decisions which directly or indirectly affect hiring, firing, pay or advancement. The validation is best done with a trial population but if time or the total size of the population preclude this, the validation must be done using results from a regular administration. The validation must be done prior to implementing any personnel decision based wholly or in part on the testing. For the validation a test group of at least 25 persons representative of the population to be tested should be used.

Criterion Referenced Tests. When tests are scored on the basis of absolute criterial rather than relative criteria refers to that process as criterion referencing. Rather than converting to standard scores or percentages on the basis of norms or relative performances of all test takers, criterion-referenced interpretation is done on the basis of the number of items answered correctly. This number may then be evaluated on the basis of a preset criterion (for example, 80 percent, 65 percent) or examined in light of the proportion of all test takers who got the item correct (called the p-value).

The principal features of criterion-referenced tests are that (1) they are constructed to measure a specific set of operationally or behaviorally stated objectives, (2) they represent an attempt to maximize content validity or appropriateness based on that set of objectives (they are often referred to as "objective-referenced"), (3) they are considered to represent samples of actual performance, and (4) performance on them can be interpreted in terms of predetermined cutoff scores.

Since most research involves comparison between groups, even criterion-referenced test results can be evaluated on a comparative basis. Tests, unless standardized, can be considered criterion-referenced.

Basis of Scoring. The items generated should be administered to a pilot group. Total scores on such measures as number of items passed should be obtained for each student in the pilot group. Performance on each item should then be compared to total scores using the following procedure, called "item analysis".

1. Subdivide students based on total score into
   a. high third scorers
   b. middle third scorers
   c. low third scorers
2. Identify the number of high third and low third scorers who pass each item (See Figure 8 for an illustration).

"Discriminability" refers to the extent to which a test item is responded to correctly by those trainees possessing less of this quality. Such items serve the purpose for which they were written--separating more knowledgeable students fromless knowledgeable ones. Only six of the ten items in Figure 9 have satisfactory discriminating power in that two-thirds (.67) or more of those who get them right are in the high third group (that is, the index of discriminability for these items is above .67). These items are 1, 3, 4, 6, 7, and 8.

"Difficulty" refers to the extent to which a test can be responded to correctly by any trainee. It differes from discriminability in that it does not include any differentiation between trainees. Of the ten items, two are too easy (items 2 and 10) since fewer than one-third (.33) got them wrong (that is, the index of difficulty for these items is below .33). Moreover, two items are too difficult (items 3 and 9) since more than two-thirds (.67) got them wrong (that is, the index of difficulty for these items is above .67). Taking both considerations into account, items 1, 4, 6, 7, and 8 would survive this pilot test since they have both satisfactory discriminability (above .67) and reasonable difficulty (between .33 and .67). Of course, you have a certain degree of flexibility in setting the cutoffs for discriminability and difficulty; in choosing to exerxise this flexibility, you should understand what these values mean and what you are doing to your test as you change them.

Notice that on those items in which a greater number of lower scores pass, the index of discriminability dips below .50 (items 9, 10). Of course, these items must be discarded. Sometimes it does not hurt to retain an extremely easy item like item 2. Since it is uniformly easy, it may have motivational value, and beceause of its uniform ease for both low and high scorers, it will not affect the relative position of scores.

Item	Number of High 1/3 Who Pass	Number of Low 1/3 Who Pass	Index of Difficulty	Index of Discriminability
1	10	0	.50	1.00
2	10	9	.05	.53
3	3	0	.85	1.00
4	9	2	.45	.82
5	5	5	.50	.50
6	10	3	.35	.77
7	7	1	.60	.87
8	9	3	.40	.75
9	2	4	.70	.33
10	8	9	.15	.47

$$\text{Index of Difficulty} = \frac{\text{number who fail an item}}{\text{total number in both groups}}$$

$$\text{For Item 1: } \frac{20-(10+0)}{10+10} = \frac{10}{20} = .50$$

$$\text{Index of Discriminability} = \frac{\text{number of high 1/3 who pass an item}}{\text{total number in both groups who pass the item}}$$

$$\text{For Item 1: } \frac{10}{10+0} = \frac{10}{10} = 1.00$$

Figure 9. Sample item analysis on the data from among a group of thirty.

In building a performance test, therefore, it is essential to: (1) outline content areas to be covered to ensure content validity; and (2) try the test out to obtain data for an item analysis in terms of item difficulty and discriminability.

It is also useful to administer any other comparable or related test to your pilot group to see whether your test relates to another test. If your test shows a relationship to another test, you can then demonstrate concurrent validity. If it relates to an aptitude test (as an example), then this may contribute to construct validity. If classroom performance records are available, additional validity tests can be made. For a performance test, however, the establishment of content validty (a nonstatistical concept) and the use of item analysis are usually sufficient for test construction.

Attemps to establish forms of validity other than content validity are usually unecessary for tests of performance, although it is useful to establish their reliability. While an item analysis contributes to the establishment of internal reliability, the Kuder-Richardson formula can also be used for this purpose.